XP-002157890

A 1200 BPS SPEECH CODER BASED ON MELP

Tian Wang, Kazuhito Koishida, Vladimir Cuperman, Allen Gersho
SignalCom, Inc.
7127 Hollister Ave. Suite 109, Goleta, California 93117, USA
E-mail: dsp@dsp-signal.com

John S. Collura
National Security Agency
9800 Savage Rd. STE 6516, Ft. Meade, MD 20755-6516, USA

ABSTRACT

This paper presents a 1.2 kbps speech coder based on the MELP analysis algorithm. In the proposed coder, the MELP parameters of three consecutive frames are grouped into a superframe and jointly quantized to obtain a high coding efficiency. The inter-frame redundancy is exploited with distinct quantization schemes for different unvoiced/voiced (U/V) frame combinations in the superframe. Novel techniques for improving performance make use of the superframe structure. These include pitch vector quantization using pitch differentials, joint quantization of pitch and U/V decisions and LSF quantization with a forward-backward interpolation method. Subjective test results indicate that the 1.2 kbps speech coder achieves approximately the same quality as the proposed federal standard 2.4 kbps MELP coder.

1. INTRODUCTION

Speech coding at 2.4 kbps and below is important for several applications, such as secure voice and satellite communications. For secure voice communications, the traditional 2.4 kbps Linear Predictive Coding (LPC) vocoders were developed 20 to 30 years ago and a particular version, LPC10, was adopted as a federal standard. In March 1996, the US government's Digital Voice Processing Consortium (DDVPC) selected the 2.4 kbps Mixed Excitation Linear Prediction (MELP) [1] speech-coding algorithm to be a new standard for narrow band secure voice coding products and applications. The MELP algorithm leads to a significant improvement in both speech quality and intelligibility compared to the LPC10 algorithm. However, in some difficult radio channels, robust 2.4 kbps transmission is not always possible. To accommodate such adverse transmission conditions, a lower bit rate is required. Previously, inter-frame redundancy was exploited with a superframe structure applied to LPC10 with an eight-frame superframe structure [4]. Recently, improvements to the modeling and quantization of the original MELP algorithm were shown to increase quality while allowing a reduced rate of 1.7 kbps [5].

In this paper, a new 1.2 kbps speech coder is proposed. The proposed coder, called Multi-Frame MELP or MF-MELP, shares the core analysis algorithm with the 2.4 kbps MELP standard, and its transmitted parameters are the same as those of the 2.4 kbps MELP coder. In the proposed coder, the MELP parameters of three consecutive frames are grouped together into a superframe and jointly quantized to obtain high coding efficiency. To take advantage of such a long frame structure, novel quantization schemes are introduced for the superframe. The proposed coding algorithm is described in this paper and the results of subjective tests are reported. These results show that the substantial redundancy that exists across three speech frames is sufficient to achieve roughly the same quality as the 2.4 kbps MELP at half the bit-rate.

2. CODER OVERVIEW

The MF-MELP coder is based on the MELP analysis algorithm. The MELP transmitted parameters are extracted every 22.5 ms frame (or 180 samples of speech at a sampling rate of 8 kHz). A superframe structure of length 67.5 ms comprising three consecutive frames is adopted in the proposed coder. The MELP parameters for each frame in the superframe are jointly quantized to obtain high coding efficiency. A pitch smoother is incorporated to avoid large pitch errors, and this results in an increase in the look-ahead by 129 samples. The total algorithmic delay for MF-MELP is 103.75 ms.

The quantization schemes of the MF-MELP are designed so that the superframe structure is efficiently exploited by using vector quantization (VQ) and interpolation. The statistical properties of voiced (V) and unvoiced (U) speech are also taken into account. Each superframe is categorized into one of several coding states with a different bit allocation for each state. State selection is done according to the U/V pattern of the superframe. Moreover, since an incorrect state identification by the decoder causes a serious degradation in the synthesized speech, the MF-MELP utilizes several techniques for reducing the effect of state mismatch between encoder and decoder due to channel errors.

3. QUANTIZATION OF PITCH AND U/V DECISIONS

3.1 Pitch Quantization

The pitch information is transmitted only for voiced frames. Different pitch quantization schemes are used for different U/V combinations in the superframe. Within those superframes where the voicing pattern contains either two or three voiced frames, the pitch parameters are vector-quantized. For voicing patterns

0-7803-6293-4/00/$10.00 ©2000 IEEE.

containing only one voiced frame, the scalar quantizer used in the MELP standard is applied for the pitch of the voiced frame. For the UUU voicing pattern, no pitch information is transmitted.

The pitch values, P_i (i=1,2,3), obtained from the pitch analysis are transformed into logarithmic values, $p_i = \log P_i$, prior to quantization. For each superframe, a pitch vector is constructed with components equal to the log pitch value for each voiced frame and a zero value for each unvoiced frame. For voicing patterns with two or three voiced frames, the pitch vector is quantized using a VQ algorithm with a new distortion measure. This distortion measure incorporates pitch differentials into the codebook search, which makes it possible to consider the time evolution of the pitch. This feature is motivated by the perceptual importance of adequately tracking the pitch trajectory.

The pitch VQ algorithm has three steps for obtaining the best index:

Step 1: Select the M-best candidates using the weighted squared Euclidean distance measure:

$$d = \sum_{i=1}^{3} w_i |p_i - \hat{p}_i|^2$$

where the weighting coefficient is defined by

$$w_i = \begin{cases} 1, & \text{for voiced frame} \\ 0, & \text{for unvoiced frame} \end{cases}$$

and p_i and $\hat{p}_i$ are the unquantized and quantized log pitch values, respectively. The above equation indicates that only voiced frames are taken into account in the codebook search.

Step 2: Calculate the differentials of the unquantized log pitch values using

$$\Delta p_i = \begin{cases} p_i - p_{i-1}, & \text{if } i-\text{th and } (i-1)-\text{th frames are voiced} \\ 0, & \text{otherwise} \end{cases}$$

for i = 1, 2, 3, where p_0 is the last log pitch value of the previous superframe. For the pitch candidates selected in step 1, calculate the quantized differentials by replacing Δp_i and p_i by $\Delta \hat{p}_i$ and $\hat{p}_i$ respectively in the equation above, where $\hat{p}_0$ is the quantized version of p_0.

Step 3: Select the optimum index from the M-best candidates that minimizes

$$d' = \sum_{i=1}^{3} w_i |p_i - \hat{p}_i|^2 + \delta \sum_{i=1}^{3} |\Delta p_i - \Delta \hat{p}_i|^2 = d + \delta \sum_{i=1}^{3} |\Delta p_i - \Delta \hat{p}_i|^2$$

where δ is a parameter to control the contribution of pitch differentials which is set to be 1 in the proposed coder.

For superframes that contain only one voiced frame, scalar quantization of the pitch is performed. The pitch value is quantized on a logarithmic scale with a 99-level uniform quantizer ranging from 20 to 160 samples. The quantizer is the same as that in the 2.4 kbps MELP standard, where the 99 levels are mapped to a 7-bit pitch codeword and the 28 unused codewords with Hamming weight 1 or 2 are reserved for error protection.

Table 2. Joint quantization of pitch and U/V decisions.

U/V patterns	3-bit CB	9-bit CB
UUU	000	The pitch value is quantized with the same 99-level uniform quantizer as the 2.4kb/s standard. The pitch value and U/V pattern are then mapped to this 9-bit codebook.
UUV		
UVU		
VUU		
VVU	001	These U/V patterns share the same codebook containing 512 codevectors of the pitch triple.
VUV	010	
UVV	100	
VVV	011	512-level codebook A
	101	512-level codebook B
	110	512-level codebook C
	111	512-level codebook D

3.2 Joint Quantization of Pitch and U/V Decisions

The U/V decisions and pitch parameters for each superframe are jointly quantized using 12 bits. The joint quantization scheme is summarized in Table 2. In this scheme, the allocation of 12-bits consists of 3 mode bits (representing the 8 possible combinations of U/V decisions for the 3 frames in a superframe) and the remaining 9 bits for pitch values. The scheme employs six separate pitch codebooks, five having 9 bits (i.e. 512 entries each) and one being the scalar quantizer; the specific codebook is determined according to the bit patterns of the 3-bit codeword representing the quantized voicing pattern. Therefore the U/V voicing pattern is first encoded into a 3-bit codeword, which is then used to select one of the 6 codebooks shown. The set of 3 pitch values is vector-quantized with the selected codebook to generate a 9-bit codeword that identifies the quantized set of 3 pitch values. Note that four codebooks are assigned to the superframes in the VVV mode, which means that the pitch vectors in the VVV-type superframes are quantized by one of 2048 codewords. If the number of voiced frames in the superframe is not larger than one, the 3-bit codeword is set to 000 and the distinction between different modes is determined within the 9-bit codebook. Note that the latter case consists of the 4 modes UUU, VUU, UVU, and UUV. In this case, the 9 available bits are more than sufficient to represent the mode information as well as the pitch value since there are 3 modes with 128 pitch values and one mode with no pitch value.

4. LSF QUANTIZATION

4.1 Quantization Procedure

Table 3 shows the bit allocation for quantizing the line spectral frequencies (LSFs). In the table, the original LSF vectors for the three frames are denoted by l_1, l_2 and l_3. For the UUU, UUV, UVU and VUU modes, the LSF vectors of unvoiced frames are quantized using a 9-bit codebook, while the LSF vector of the

Table 2. Bit allocation for LSF quantization.

U/V pattern	LSF l_1	LSF l_2	LSF l_3	Inter. Coef.	Res. of l_1, l_2	Total
UUU	9	9	9	0	0	27
VUU	7666	9	9	0	0	43
UVU	9	7666	9	0	0	43
UUV	9	9	7666	0	0	43
UVV VUV VVV	8	0	7666	4	86	43
VVU	8	0	9	4	8666	39

Table 3. Bit allocation of 1.2 kbps MF-MELP coder for a superframe of 67.5 ms.

Parameters	U/V patterns of superframe				
	VVV	UVV VUV	VVU	UUV UVU VUU	UUU
Pitch & Global UV Decisions	12	12	12	12	12
LSFs	43	43	39	43	27
Gains	10	10	10	10	10
Bandpass Voicing	6	4	4	2	0
Fourier Magnitudes	8	8	8	8	0
Aperiodic Flag	1	1	1	1	0
Synchronization	1	1	1	1	1
Error Protection	0	2	6	4	31
Total	81	81	81	81	81

voiced frame is quantized with the same 25-bit multi-stage VQ (MSVQ) quantizer as in the MELP standard.

The LSF vectors for the other U/V patterns are encoded using a forward-backward interpolation scheme. This scheme works as follows. First the LSFs of the last frame in the current superframe, l_3, are quantized to $\hat{l}_3$ using the 9-bit codebook for unvoiced case or the same 25-bit MSVQ codebook as in the MELP coder for voiced case. Predicted values of l_1 and l_2 are then obtained by interpolating $\hat{l}_p$ and $\hat{l}_3$ as follows ($\hat{l}_p$ is the quantized LSFs of the last frame of the previous superframe):

$$\tilde{l}_1(j) = a_1(j)\hat{l}_p(j) + [1 - a_1(j)]\hat{l}_3(j)$$
$$\tilde{l}_2(j) = a_2(j)\hat{l}_p(j) + [1 - a_2(j)]\hat{l}_3(j) \quad j = 1, \cdots, 10$$

where $a_1(j)$ and $a_2(j)$ are the interpolation coefficients, and $\hat{l}_i(j)$ is the j-th component of $\hat{l}_i$. The coefficients are stored in a codebook and the best set of the coefficients are selected by minimizing the distortion measure:

$$E = \sum_{j=1}^{10} w_1(j)\left|l_1(j) - \tilde{l}_1(j)\right|^2 + \sum_{j=1}^{10} w_2(j)\left|l_2(j) - \tilde{l}_2(j)\right|^2$$

where $w_i(j)$ are the weighting coefficients obtained with the same procedure as in the 2.4 kbps MELP standard. After obtaining the best interpolation coefficients, the residual LSF vector for frames 1 and 2 are computed by

$$r_1(j) = l_1(j) - \tilde{l}_1(j)$$
$$r_2(j) = l_2(j) - \tilde{l}_2(j) \qquad j = 1, \cdots, 10.$$

The two residual vectors are concatenated and the resulting 20-dimension residual vector is encoded with a MSVQ quantizer.

4.2 Design Method for Interpolation Codebook

The interpolation codebook is designed using the generalized Lloyd algorithm. In the training algorithm, two procedures are alternately performed. The first procedure encodes LSFs vectors of training database using the distortion measure E. The second procedure optimizes the interpolation codebook in such a way that the summation of all the superframe distortions is minimized. By setting the partial derivatives of the summation of the distortions with respect to $a_1(j)$ and $a_2(j)$ to zero, the optimum interpolation coefficients are obtained from

$$a_1(j) = \frac{\sum_n w_{1,n}(j)\left[\hat{l}_{3,n}(j) - l_{1,n}(j)\right]\left[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)\right]}{\sum_n w_{1,n}(j)\left[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)\right]^2}$$

$$a_2(j) = \frac{\sum_n w_{2,n}(j)\left[\hat{l}_{3,n}(j) - l_{2,n}(j)\right]\left[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)\right]}{\sum_n w_{2,n}(j)\left[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)\right]^2}$$

for j=1,...,10. The interpolation coefficients codebook was trained and tested for several codebook sizes. A codebook with 16 entries was found to be quite efficient.

5. BIT ALLOCATION

The bit allocation of the 1.2 kb/s coder is summarized in Table 3. In the coder, two gain parameters are calculated per frame, with 6 gains per superframe. The 6 gain parameters are vector-quantized in logarithmic domain using a 10-bit codebook.

The binary voicing decisions for 5 bands are obtained per frame. The bandpass information for the lowest band is determined from the U/V decision. The bandpass decisions of the remaining 4 bands are employed only for voiced frames and quantized with a 2-bit codebook.

The Fourier magnitude vector is computed only for voiced frames. The vector of the last voiced frame in the current superframe is quantized with the same 8-bit quantizer as the MELP standard. The Fourier magnitude vectors for the other

Table 4. DRT testing results.

Description	DRT Scores		
	Quiet	HMMWV	0.5% BER
2.4 kbps MELP	93.8	72.5	93.4
1.2 kbps MF-MELP	92.1	70.2	86.3

voiced frames are reconstructed using the quantized vectors of the current and previous superframes. The reconstruction procedure uses either an interpolation or a repetition method according to the U/V decisions.

The aperiodic flag is also obtained only from voiced frames. The MF-MELP coder uses 1-bit per superframe for the quantization of the aperiodic flag. Different 1-bit codebooks are employed for different U/V patterns.

6. TEST RESULTS

Two subjective tests were conducted to evaluate the performance of the MF-MELP speech coder. One is the Diagnostic Rhyme Test (DRT) for measuring speech intelligibility, and another is the Diagnostic Acceptability Measure (DAM) for speech quality. For a comparison purpose, the 2.4 kbps MELP standard was included in both tests. The 1.2 and 2.4 kbps coders were tested in quiet background, HMMWV noise environment and 0.5% random bit error channel. Note that the HMMWV is a heavy-duty four-wheeled drive vehicle used for troop transport. The noise pre-processor proposed in Ref. [3] is integrated into the 1.2 and 2.4 kbps coders.

The DRT and DAM test results are shown in Table 4 and 5, respectively. It is shown that, in quiet and noise environments, the MF-MELP coder provides comparable intelligibility to the 2.4 kbps MELP coder. Although the proposed coder obtains lower DAM scores than the MELP standard, the quality of the 1.2 kb/s coder is reasonably close to that of the 2.4 kbps MELP coder.

For the channel error condition, the difference in intelligibility score between the 1.2 and 2.4 kbps coders is larger than that in other conditions. An efficient way for increasing the robustness against transmission errors is to use a parity check bit for the 3-bit mode codebook in Table 1. Another version of the MF-MELP coder with the parity check was designed, in which, to save 1-bit for the parity check, the voiced LSF parameters are quantized with a 24-bit MSVQ codebook instead of the 25-bit MSVQ codebook. The MF-MELP coder with the parity check is found to improve the score by 3.7 in the channel error condition, while the score for clean channel (quiet background) degrades by 1.8.

7. CONCLUSIONS

This paper has introduced a 1.2 kbps speech coder, MF-MELP, based on the MELP analysis algorithm. In the proposed coder, the MELP transmitted parameters of consecutive three frames are quantized together. Efficient vector quantization schemes are employed depending on the different U/V decisions for the superframe, taking into account statistical properties of voiced

Table 5. DAM testing results.

Description	DAM Scores	
	Quiet	HMMWV
2.4 kbps MELP	68.2	52.2
1.2 kbps MF-MELP	61.8	48.9

and unvoiced speech. The MF-MELP coder incorporates novel techniques for improving the performance, such as pitch quantization with pitch differentials, joint quantization of pitch and U/V decisions and LSF quantization with a forward-backward interpolation method. The DAM and DRT test results have indicated that the MF-MELP coder has approximately the same speech quality as the 2.4 kbps MELP standard.

8. REFERENCES

[1] A.V. McCree, T.P. and Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 4, pp. 242-250, July 1995.

[2] L.M. Supplee, R.P. Chon, J.S. Collura and A.V. McCree, "MELP: The New Federal Standard at 2400 bps," in Proc. ICASSP-97, vol. 2, pp. 1591-1594, 1997.

[3] R. Martin and R.V. Cox, "New Speech Enhancement Techniques for Low Bit Rate Speech Coding," in Proc. Speech Coding Workshop-99, pp. 165-167, 1999.

[4] D.P. Kemp, J.S. Collura, T.E. Tremain, "Multi-frame coding of LPC parameters at 600-800 bps" Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, vol.1, pp. 609-612, 1991.

[5] A.V. McCree and J.C. De Martin, "A 1.7 kb/s MELP coder with improved analysis and quantization" Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, pp. 593-596, 1998.